ARTICLE

Rare, Low-Frequency, and Common Variants in the Protein-Coding Sequence of Biological Candidate Genes from GWASs Contribute to Risk of Rheumatoid Arthritis

Dorothée Diogo,[1,2,3,19] Fina Kurreeman,[1,2,3,13,19] Eli A. Stahl,[1,2,3] Katherine P. Liao,[1] Namrata Gupta,[4] Jeffrey D. Greenberg,[5] Manuel A. Rivas,[3] Brendan Hickey,[1] Jason Flannick,[3,6] Brian Thomson,[3] Candace Guiducci,[3] Stephan Ripke,[3,7,8] Ivan Adzhubey,[2] Anne Barton,[9] Joel M. Kremer,[10] Lars Alfredsson,[11] Consortium of Rheumatology Researchers of North America, Rheumatoid Arthritis Consortium International, Shamil Sunyaev,[2,3] Javier Martin,[12] Alexandra Zhernakova,[13,14] John Bowes,[9] Steve Eyre,[9] Katherine A. Siminovitch,[15,16] Peter K. Gregersen,[17] Jane Worthington,[9] Lars Klareskog,[18] Leonid Padyukov,[18] Soumya Raychaudhuri,[1,2,3,9] and Robert M. Plenge[1,2,3,*]

The extent to which variants in the protein-coding sequence of genes contribute to risk of rheumatoid arthritis (RA) is unknown. In this study, we addressed this issue by deep exon sequencing and large-scale genotyping of 25 biological candidate genes located within RA risk loci discovered by genome-wide association studies (GWASs). First, we assessed the contribution of rare coding variants in the 25 genes to the risk of RA in a pooled sequencing study of 500 RA cases and 650 controls of European ancestry. We observed an accumulation of rare nonsynonymous variants exclusive to RA cases in *IL2RA* and *IL2RB* (burden test: $p = 0.007$ and $p = 0.018$, respectively). Next, we assessed the aggregate contribution of low-frequency and common coding variants to the risk of RA by dense genotyping of the 25 gene loci in 10,609 RA cases and 35,605 controls. We observed a strong enrichment of coding variants with a nominal signal of association with RA ($p < 0.05$) after adjusting for the best signal of association at the loci ($p_{enrichment} = 6.4 \times 10^{-4}$). For one locus containing *CD2*, we found that a missense variant, rs699738 (c.798C>A [p.His266Gln]), and a noncoding variant, rs624988, reside on distinct haplotypes and independently contribute to the risk of RA ($p = 4.6 \times 10^{-6}$). Overall, our results indicate that variants (distributed across the allele-frequency spectrum) within the protein-coding portion of a subset of biological candidate genes identified by GWASs contribute to the risk of RA. Further, we have demonstrated that very large sample sizes will be required for comprehensively identifying the independent alleles contributing to the missing heritability of RA.

Introduction

Genome-wide association studies (GWASs) have successfully identified many loci that influence a wide variety of complex diseases. However, a large portion of the heritability of complex traits has not been explained by GWASs.[1] Several hypotheses have been proposed to explain the missing heritability from association studies. One hypothesis involves rare and low-frequency variants, which are not well captured by current genotyping arrays. These variants are expected to be under purifying selection and thus enriched with deleterious, protein-coding mutations participating in complex traits.[2,3] Another hypothesis to explain the missing heritability involves common variants with small effect sizes; these variants do not reach genome-wide significance thresholds in GWASs.

Methods have been developed for the assessment of the genetic architecture of complex traits with the use of GWAS data.[4,5] In a recent study,[6] we used polygenic models to demonstrate that common variants with weak effect sizes account for a large portion of the missing genetic contribution to diseases. Furthermore, we conducted Bayesian inference analyses and found that a portion of the underlying risk is contributed by low-frequency and rare causal variants and that these variants have modest effect sizes (odds ratios [ORs] < 1.5). Our simulations suggested that at least some disease risk loci harbor multiple independent risk alleles and that both common variants

[1]Division of Rheumatology, Immunology, and Allergy, Brigham and Women's Hospital, Harvard Medical School, Boston, MA 02115, USA; [2]Division of Genetics, Brigham and Women's Hospital, Harvard Medical School, Boston, MA 02115, USA; [3]Medical and Population Genetics Program, Broad Institute, Cambridge, MA 02142, USA; [4]Genomics Platform at Broad Institute, Cambridge, MA 02142, USA; [5]New York University Hospital for Joint Diseases, New York, NY 10003, USA; [6]Department of Molecular Biology and Diabetes Unit, Massachusetts General Hospital, Boston, MA 02114, USA; [7]Analytic and Translational Genetics Unit, Massachusetts General Hospital and Harvard Medical School, Boston, MA 02114, USA; [8]Stanley Center for Psychiatric Research, Broad Institute of MIT and Harvard, Cambridge, MA 02142, USA; [9]Arthritis Research UK Epidemiology Unit, University of Manchester, Manchester Academic Health Sciences Centre, Manchester M13 9PT, UK; [10]The Center of Rheumatology, Albany, NY 12206, USA; [11]Institute of Environmental Medicine, Karolinska Institutet, 171 77 Stockholm, Sweden; [12]Instituto de Parasitología y Biomedicina López-Neyra, Consejo Superior de Investigaciones Científicas, 18100 Granada, Spain; [13]Department of Rheumatology, Leiden University Medical Centre, 2300 RC Leiden, The Netherlands; [14]Genetics Department, University Medical Center and Groningen University, 9700 RB Groningen, The Netherlands; [15]Samuel Lunenfeld Research Institute, Mount Sinai Hospital, Toronto, ON M5G 1X5, Canada; [16]Department of Medicine, University of Toronto, Toronto, ON M5G 1X5, Canada; [17]The Feinstein Institute for Medical Research, North Shore-Long Island Jewish Health System, Manhasset, NY 11030, USA; [18]Rheumatology Unit, Department of Medicine, Karolinska Institutet and Karolinska University Hospital Solna, 171 76 Stockholm, Sweden
[19]These authors contributed equally to this work
*Correspondence: rplenge@partners.org
http://dx.doi.org/10.1016/j.ajhg.2012.11.012.

and independent low-frequency or rare causal variants contribute to disease risk.

Investigating independent risk alleles affecting protein-coding regions in associated loci appears to be a good strategy for identifying genes of biological relevance in complex traits. Indeed, several examples in the literature show that genes containing common variants associated with complex traits also contain rare and low-frequency protein-coding variants; such genes include *PCSK9* (MIM 607786; associated with low-density-lipoprotein cholesterol levels),[7] *IFIH1* (MIM 606951; associated with type 1 diabetes),[8] genes associated with hypertriglyceridemia[9] or inflammatory bowel disease,[10,11] *CFH* (MIM 134370; associated with age-related macular degeneration),[12] *MTNR1B* (MIM 600804; associated with type 2 diabetes),[13] *SHANK2* (MIM 603290; associated with autism spectrum disorders),[14] and *CARD14* (MIM 607211; associated with psoriasis).[15]

Here, we aimed to further assess the role of rare, low-frequency, and common variants with weak effects on the genetic architecture of rheumatoid arthritis (RA [MIM 180300]). We focused on variants within protein-coding regions (e.g., missense, nonsense, and synonymous variants) because it is more straightforward to annotate biological function and because independent protein-coding variants can help pinpoint causative genes. Our findings support our simulated genetic models and provide strong evidence that rare, low-frequency, and common variants within protein-coding sequences of biological candidate genes from GWASs contribute to the risk of RA.

Subjects and Methods

Samples

Our sequencing study included 500 RA cases and 650 matched controls of European ancestry. RA cases were selected on the basis of a high titer of anticitrullinated protein antibodies (ACPAs), markers of disease severity.[16] These samples originated from two different collections. A total of 250 RA cases and 250 controls were recruited from Sweden as part of the Epidemiological Investigation of Rheumatoid Arthritis.[17] The remaining 250 cases and 400 controls were recruited from the United States as part of a study using electronic medical records.[18] Blood samples were collected according to protocols approved by local institutional review boards. All individuals provided informed consent.

Exon Sequencing

We targeted 25 biological candidate genes in RA-associated loci by using GRAIL[19] for exon resequencing. We combined DNA in 10 pools of RA cases and 13 pools of matched controls, and each pool contained the same amount of DNA from 50 individuals.

We matched case and control samples in pools for sequencing by first calculating principal-component (PC) distances between all pairs of samples as Euclidean distances along five eigenvalue-weighted PCs (calculated from GWAS data). We matched one control sample to each case by randomly choosing from nearby controls (probability was inversely proportional to PC distance) and minimizing the total case-control PC distance over 100 iterations, and we excluded outliers from the distribution of case-control PC distances. We then established case pools by randomly choosing pools from nearby cases and minimizing total within-pool PC distance over 1,000 iterations, and matched controls constituted matching control pools. For each pool, we performed PCR amplification to capture the target sequence. We then combined all PCR amplicons (~125 bp per amplicon) in equimolar concentrations. Each pool was paired-end sequenced at the Broad Institute on one lane of the Illumina Genome Analyzer II. Reads of 125 bp were aligned to the reference human genome (NCBI Build 36/hg18) with the MAQ algorithm[20] within the Picard analysis pipeline, similar to methods described in other studies.[11] We used the method Syzygy to call variants on the pooled sequencing data.[11] We applied several filters to identify high-quality variants in each pool. First, we considered only the positions with $\geq$2,000× coverage, i.e., a minimum of 20× coverage per chromosome. Second, we required concordant allele frequencies on the forward and reverse strands. Third, we considered the nonrandomness of the noise spectrum of technical artifacts due to a biased preference for different base signal channels. Fourth, we filtered out all SNPs that clustered together within a 5 bp window centered on a SNP. Finally, because we sequenced the 23 pools in three separate batches, we performed regression analyses to determine whether significant batch effects existed in our data. After these stringent filtering criteria, 281 coding variants were called.

We used GWAS data available for 250 RA cases and 250 controls to assess the quality of the variants called. First, we targeted 18 low-frequency variants that were genotyped with GWAS arrays and determined whether we were able to detect singletons, doubletons, tripletons, and all alleles present at a frequency $\geq$ 4 in each pool sequenced. We detected 99% of all singletons and 100% of all doubletons and tripletons in our sequencing data. Our approach missed one singleton as a result of low coverage at that base ($n_{reads} = 184$). To determine the specificity in our data, we took advantage of a larger set of SNPs present in both our sequencing data and our GWAS data (n = 40, genotyped or imputed). The allele frequencies estimated from read counts correlated strongly with expected frequencies in the GWAS pools (pearson correlation, $R^2_{cases} = 0.990$ and $R^2_{controls} = 0.999$). We achieved slightly diminished specificity ($R^2_{cases} = 0.934$ and $R^2_{controls} = 0.972$) with lower-frequency variants ($frequency_{syzygy} < 0.05$, n = 25; Table S2 and Figure S1, available online). We used PolyPhen[21] to annotate the 281 variants (missense, nonsense, or synonymous) and predict their impact on the structure and function of the protein.

Analysis of Burden Association Signal Due to Rare Variants

To assess the overall genetic burden due to rare variants in the genes investigated, we used four collapsing methods described in the literature: (1) the weighted-sum statistic described by Madsen and Browning,[22] (2) the variable-threshold model described by Price et al.,[23] (3) the T1 and T5 models by Morris and Zeggini,[24] and (4) the C-alpha test described by Neale et al.[25] In this analysis, we only included variants that were not previously described in the 1000 Genomes CEU (Utah residents with ancestry from northern and western Europe from the CEPH collection) panel (release June 2011) or in dbSNP. For each gene, we separated the nonsynonymous and synonymous variants. We then pooled the variants on the basis of their minor allele frequency (MAF) in controls (MAF < 1% or MAF < 5%) and tested their cumulative

effect by using the four collapsing methods and allele counts in the pooled data (note: singletons were not included in the C-alpha test). We repeated the analysis by incorporating PolyPhen predictions in the statistical tests and giving higher weight to variants predicted to be functionally relevant. We assessed statistical significance by 100,000 case-control permutations.

Single SNP Analysis Using Genotype Data

To test the low-frequency and common coding variants for association with RA, we used two types of genotyping data available in our laboratory (Table S3). Seven case-control collections were genotyped with the Illumina Immunochip platform as part of the Rheumatoid Arthritis Consortium International,[26] the i2b2 program,[18] or the Consortium of Rheumatology Researchers of North America (CORRONA).[27] Four additional case-control collections from our previous GWAS were included.[28]

The i2b2 samples were identified with the use of clinical data from the electronic medical records (EMRs) as previously described.[29] The i2b2 and CORRONA samples were genotyped together at the Broad Institute. Genotype calling was performed on all samples as a single project with the GenomeStudio Data Analysis Software package. Initial genotype clustering was performed with the default Illumina cluster file (Immunochip_Gentrain_June2010.egt) and manifest file Immuno_BeadChip_11419691_B.csv. Extensive quality control and data filtering were performed as described elsewhere,[26] and SNPs with $p < 10^{-2}$ in a chi-square test for difference in missingness between cases and controls were also removed, leaving 148,972 SNPs for subsequent analysis. We performed PC analysis by using EIGENSOFT[30] with HapMap phase III samples to exclude individuals of non-European ancestry.

We calculated identity-by-state estimates by using PLINK[31] to remove related samples across the Immunochip and GWAS collections with the use of a set of SNPs with a missing-genotype rate < 0.5%, MAF > 5%, and pruned linkage disequilibrium (LD). The 11 collections resulted in a total sample size of 10,609 RA ACPA$^+$ cases and 35,605 controls of European ancestry. We applied stringent filters on each Immunochip data set to select only high-quality SNPs for association testing (call rate > 0.99 in cases and controls, MAF > 0.1%, Hardy-Weinberg equilibrium $p < 5 \times 10^{-7}$, and $p > 10^{-2}$ in chi-square test for difference in missingness between cases and controls). GWAS collections were imputed with the use of SNPs from the 1000 Genomes CEU reference panel (release June 2011), and SNPs with bad statistical information (info score < 0.4) or a MAF < 0.1% were removed in the subsequent analysis.

To test for association with RA risk, we used PLINK to conduct logistic-regression analyses of the 11 RA case-control collections, and this included ten PCs calculated as covariates with EIGENSOFT.[30] After checking genomic-control inflation in each collection, we conducted an inverse-variance-weighted meta-analysis to combine the results across the 11 collections at each of the loci of interest. We also computed Cochran's Q statistics and I^2 statistics to assess heterogeneity across collections. Meta-analysis and computation of heterogeneity statistics were adapted from the MANTEL program.[32] At each of the loci analyzed, we performed a conditional analysis in PLINK to test for an independent signal of association by adjusting for the genotypes at the best signal of association at the locus. When the SNP with the best signal of association was not present in the 11 collections, the best proxy ($r^2 > 0.9$) present in the 11 collections was used in the conditional analysis. We also performed a conditional analysis by adjusting for the coding variants to assess the contribution of these coding variants to the known GWAS signals.

To assess the enrichment of coding variants with nominal signal of association in our meta-analysis and conditional analysis ($p_{observed} < 0.05$), we generated 1,000 sets of permuted phenotypes for each of the 11 collections. For each collection, we performed 1,000 logistic regressions (including ten PCs as covariates) by using the permuted phenotypes. We then performed 1,000 meta-analyses of the logistic-regression results from the 11 collections. We extracted the p values ($p_{permutation}$) for each of the SNPs with $p_{observed} < 0.05$ in our initial meta-analysis or conditional analysis. For several p value thresholds ($p_{threshold}$ ranging from 5×10^{-8} to 1), we compared the number of SNPs with $p_{observed} < p_{threshold}$ and $p_{permutation} < p_{threshold}$ and assessed the significance of the results by using Fisher's exact tests.

To assess the significance of observing coding SNPs in LD ($r^2 > 0.7$) with the best signals of association at the loci, we randomly extracted 1,000 SNPs from each of the loci of interest. We then compared the frequency of randomly extracted SNPs that showed $r^2 > 0.7$ with the best hit in the conditional analysis to the observed frequency among coding SNPs with $p < 0.05$ in the conditional analysis (Fisher's exact test).

Independent Signal of Association at *CD2*

To investigate the independent signal of association at *CD2*, we performed a conditional haplotype analysis in PLINK by using only samples genotyped in the Immunochip (7,222 RA cases and 15,870 controls) and including PCs as covariates. To assess the significance of the results controlling for the known common variant, we permuted case-control status while preserving genotypes for the common variant (fixing case-control allele frequencies and ORs). In 5,000 permutations, we evaluated the frequency of observing a signal of association at $p \leq 0.015$ when controlling for the common variant.

To investigate the best causal candidate variant responsible for this independent signal, we extracted all SNPs that were described in the 1000 Genomes CEU data (release June 2011) and that were in strong LD ($r^2 > 0.8$) with rs798036. We annotated the 17 proxies by using PolyPhen,[21] SIFT,[33] GERP,[34] and publically available data on expression quantitative trait loci (eQTL).

Power Calculations

To assess the number of samples required for reaching significant p values in our burden tests, we extrapolated our results in *IL2RA* and *IL2RB*, as described previously.[35] We used (1) ORs estimated on the basis of singleton counts in our 500 RA cases and in European American samples as part of the National Heart, Lung, and Blood Institute (NHLBI) Exome Sequencing Project (ESP) and (2) ORs estimated on the basis of allele counts for variants with a MAF < 5% in our RA cases and controls. We calculated burden p values by using a one-sided Fisher's exact test.

Power to observe a $p < 5 \times 10^{-8}$ for association at the *CD2* common missense variant was assessed with the Genetic Power Calculator.[36]

Results

Study Overview

An overview of our study is shown in Figure 1. We sequenced the coding exons of 25 genes located within

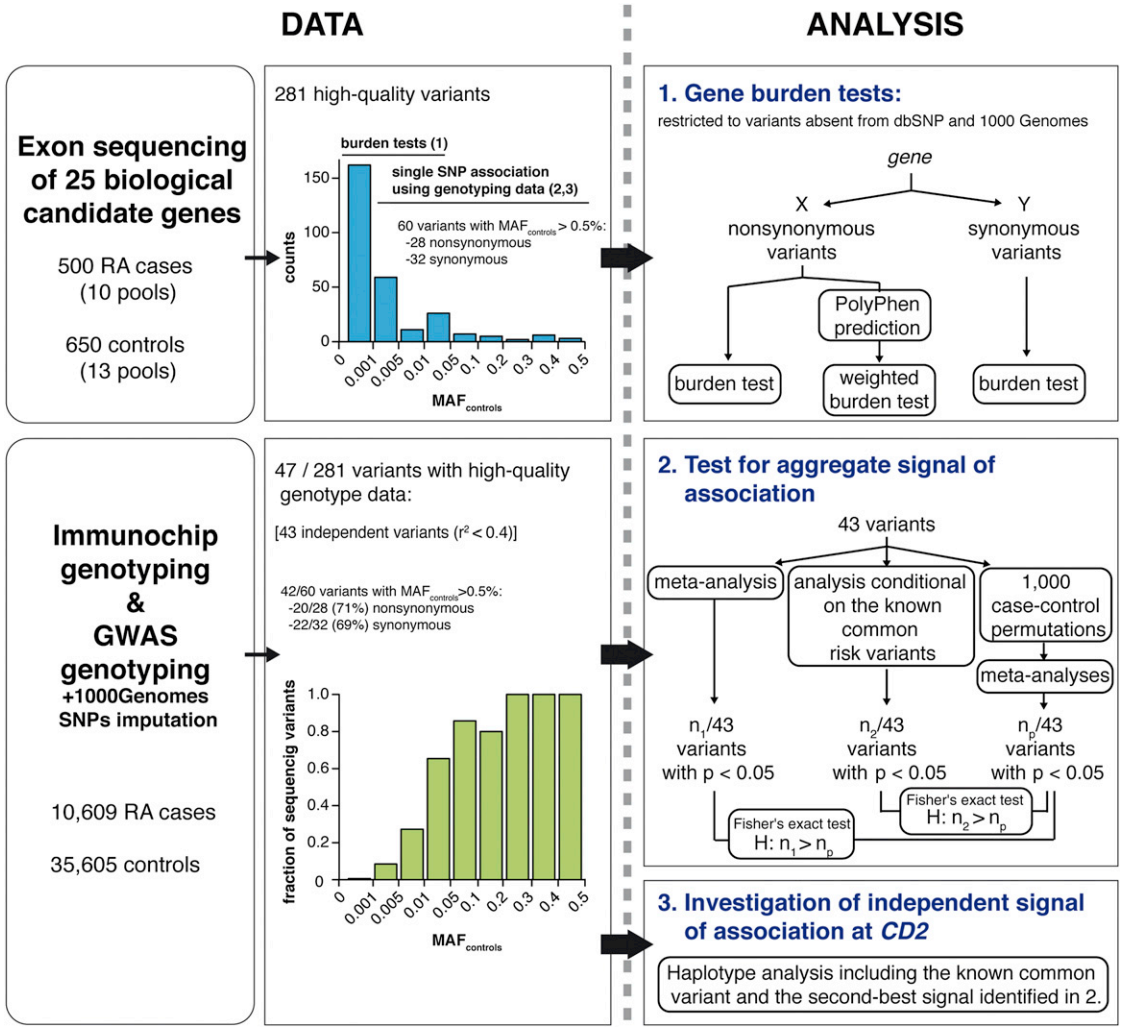


Figure 1. Description of the Study Design
Our study used two sources of data: (1) we sequenced the coding exons of 25 genes located within RA risk loci identified by GWASs, leading to the identification of 281 protein-coding variants (top panel: distribution shown for MAF in controls); and (2) we used integrated Immunochip and GWAS data for 10,609 seropositive RA cases and 35,605 controls and focused only on the protein-coding variants from these same 25 genes. We performed three types of analyses: (1) To test for association, we investigated burden association signals driven by an accumulation of rare variants (frequency < 0.5%). (2) We assessed the role of low-frequency (0.5%–5%) and common (frequency > 5%) variants with weak effect in RA by adjusting for the common SNP identified by GWASs. (3) For detailed analysis, we selected the *CD2* locus, which showed suggestive evidence of an independent signal of association in the conditional analysis.

RA risk loci identified by GWASs. We focused our analysis on the best biological candidate genes at each locus because most RA-associated loci have evidenced connectivity that implicates specific biological pathways.[19,37] We investigated the implication of variants previously not associated with risk of RA in two ways. First, we investigated burden association signals driven by an accumulation of rare variants (frequency < 0.5%). Second, we assessed the role of independent low-frequency (0.5%–5%) and common (frequency > 5%) variants with a weak effect in RA by adjusting for the known common signals of association identified by GWASs. To perform this analysis, we took advantage of a large genotyping data set including Immunochip and GWAS data. Finally, for detailed analysis, we selected the *CD2* locus, which showed suggestive evidence of an independent signal of association in the conditional analysis.

Identification of Coding Variants in Biological Candidate Genes

We selected 25 biological candidate genes from loci associated with RA in previous GWASs[28,38] by using GRAIL[19] and targeted them for exon sequencing in 500 RA cases seropositive for ACPAs and 650 matched controls of European ancestry. Overall, 86.4% (ranging between 62.6% in *FCGR2A* (MIM 146790) and 100% in *CCL21* [MIM 602737] and *CTLA4* [MIM 123890]) of the 36.8 kb target regions were sequenced with >20× coverage per chromosome in at least 80% of case pools and control pools (Table 1). The average coverage per position sequenced

Table 1. Number of Coding Variants Identified in the Biological Candidate Genes through Exon Sequencing

Gene	Coding-Sequence Length (bp)	Percentage Covered[a]	Number of Variants				
			Total	Nonsynonymous		Synonymous	
				All	Tested[b]	All	Tested[b]
BLK	1,515	78.3%	17	11	8	6	3
CCL21	372	100.0%	2	1	1	1	0
CCR6	1,122	97.9%	9	3	1	6	3
CD2	1,053	92.9%	9	8	5	1	1
CD28	660	77.3%	4	2	2	2	0
CD40	831	84.7%	4	2	1	2	1
CD58	750	80.3%	6	4	3	2	1
CTLA4	669	100.0%	2	2	1	0	0
FCGR2A	951	62.6%	7	3	2	4	2
IL2	459	86.9%	1	0	0	1	0
IL21	486	78.4%	2	0	0	2	1
IL2RA	816	92.2%	11	7	4	4	2
IL2RB	1,653	91.9%	13	9	6	4	3
IRF5	1,494	85.5%	7	4	4	3	1
PRDM1	2,475	85.9%	21	13	6	8	5
PRKCQ	2,118	97.5%	23	11	8	12	8
PTPN22	2,421	86.8%	19	14	9	5	4
PTPRC	3,912	84.6%	37	21	16	16	10
REL	1,857	77.2%	5	2	2	3	2
STAT4	2,244	91.0%	8	4	2	4	3
TAGAP	2,193	78.6%	16	10	7	6	4
TNFAIP3	2,370	97.2%	14	10	8	4	2
TNFRSF14	849	80.4%	11	6	3	5	2
TRAF1	1,248	86.1%	19	10	7	9	6
TRAF6	1,566	81.9%	14	10	10	4	4
Total	**36,084**	**86.4%**	**281**	**167**	**116**	**114**	**68**

[a]Percentage of coding sites sequenced with >20× coverage per chromosome in at least 80% of case pools and control pools.
[b]Variants absent from dbSNP and the 1000 Genomes Project (release June 2011) and included in the collapsing tests and burden tests.

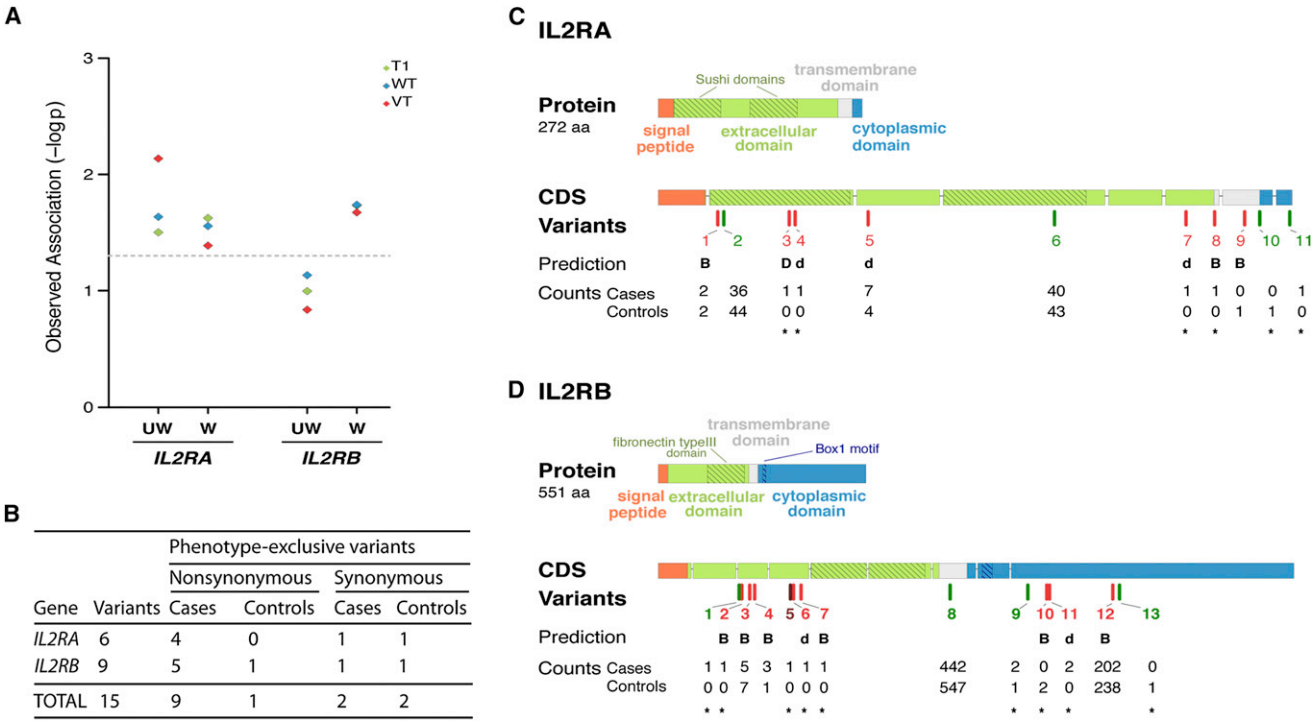
and per pool was 90,700×, i.e., 907× coverage per position per chromosome.

We used the calling variant method Syzygy[11] and identified 281 high-quality coding variants, i.e., one variant every 128 bp (Table 1 and Table S1). Using GWAS data available for 250 case-control pairs included in our sequencing study, we found high sensitivity and specificity of our sequencing results (Table S2 and Figure S1). The transition-to-transversion ratio based on the 281 SNPs was 3.6, consistent with previously published ratios from exon-sequencing data.[3]

Out of the 281 variants, 232 (83%) had a MAF less than 1% in controls, and as expected, only 97 (35%) were previously described in the 1000 Genomes Project or dbSNP, showing that most of the variants identified here are both rare and previously uncharacterized. The 281 variants included 114 synonymous, 164 missense, and 3 nonsense variants. We used PolyPhen[21] to predict the function of the missense variants. A total of 32 (20%) and 43 (26%) variants were predicted to be potentially damaging and probably damaging, respectively (Table S1). Among the synonymous variants, 26 (23%) had a conservation score > 2 as determined by GERP,[34] giving some evidence of evolutionary constrained sites (Table S1).

Accumulation of Nonsynonymous Rare Variants in *IL2RA* and *IL2RB* in Individuals with RA

To assess association at rare variants, we used four collapsing (or burden) methods described in the literature. The weighted-sum statistic,[22] the variable-threshold



B

		Phenotype-exclusive variants			
		Nonsynonymous		Synonymous	
Gene	Variants	Cases	Controls	Cases	Controls
IL2RA	6	4	0	1	1
IL2RB	9	5	1	1	1
TOTAL	15	9	1	2	2

Figure 2. Accumulation of Coding Rare Variants in *IL2RA* and *IL2RB*

(A) Burden association signal driven by nonsynonymous variants. Two types of tests were performed: unweighted (UW) tests and tests weighted (W) with PolyPhen scores. For these two genes, we did not obtain any result by using the C-alpha method because it did not include singletons.

(B) Accumulation of rare variants exclusive to RA cases in *IL2RA* and *IL2RB*.

(C and D) Distribution of variants across *IL2RA* and *IL2RB*. Missense, nonsense, and synonymous variants are shown in red, brown, and green, respectively. For the missense variants, PolyPhen prediction is indicated (B, benign; d, potentially damaging; and D, probably damaging). Variants included in the collapsing tests and burden tests (i.e., variants not described in the 1000 Genomes Project or in dbSNP) are highlighted with a star.

model,[23] and the T1 and T5 models[24] assume that the effects of the combined variants on the phenotype are in the same direction, whereas the C-alpha test[25] allows detection of opposite effects. For our primary analysis, we restricted the burden tests to variants not described in dbSNP or the 1000 Genomes Project (release June 2011) under the hypothesis that truly private mutations are more likely to be pathogenic. We tested groups of nonsynonymous ($n = 116$) and synonymous ($n = 68$) variants separately; for the nonsynonymous variants, we performed both unweighted tests and tests incorporating PolyPhen predictions of the functional effect of the variants in the statistical tests such that higher weight was attributed to variants predicted to have functional impact. In a second analysis, we performed gene burden tests by using all nonsynonymous variants identified, weighted, and unweighted with PolyPhen prediction scores.

Two genes, *IL2RA* (MIM 147730) and *IL2RB* (MIM 146710), showed nominal burden signal of association ($p < 0.05$) driven by two or more nonsynonymous variants (Figure 2A). In *IL2RA* and *IL2RB*, four and six nonsynonymous variants, respectively, were included in the tests. In *IL2RA*, incorporating PolyPhen prediction did not affect the burden signal observed, whereas *IL2RB* variants predicted to be potentially or probably damaging did influence the genetic-burden signal (Figure 2A). The same results were observed when we included all nonsynonymous variants in the tests; the only exception was *PTPN22* (MIM 600716), which only reached $p < 0.05$ when we included the known RA-risk missense variant rs2476601 (data not shown).

The burden signal at *IL2RA* and *IL2RB* corresponded to an accumulation of nonsynonymous rare variants, mainly singletons, exclusively identified in cases (Figures 2B–2D). The burden signal of association at these two genes could not be attributed to a bias in the sequencing coverage between case and control pools given that all ten variants included in the *IL2R* burden tests showed >50,000× coverage in each pool (i.e > 500× coverage per chromosome per pool). The distribution of the variants across the genes highlighted that eight out of the nine cases-exclusive nonsynonymous variants in *IL2RA* and *IL2RB* lie within the extracellular-domain-coding regions (Figures 2C and 2D).

We further assessed the frequency of singleton missense SNPs in *IL2RA* and *IL2RB* among 4,300 individuals of European ancestry from the NHLBI ESP. None of the singleton variants in *IL2RA* and *IL2RB* were present in the ESP data (four *IL2RA* and four *IL2RB* singleton variants in RA cases). Among ESP samples, only six missense singletons

(12 missense SNPs with a MAF < 5%) were identified in *IL2RA* and 12 missense singletons (22 missense SNPs with a MAF < 5%) were identified in *IL2RB*. Using our sequencing results from RA samples (n = 500 RA cases) and the ESP data as controls (n = 4,300 controls) in one-sided Fisher's exact tests incorporating singletons, we observed the following ORs: *IL2RA*, OR = 5.75 (p = 0.02); and *IL2RB*, OR = 2.9 (p = 0.08). These results are consistent with the observed effect sizes from our sequencing results in *IL2RB* (OR = 3.0), providing supporting evidence that missense mutations in these two genes contribute to risk of RA. (We could not determine the OR in *IL2RA* on the basis of our sequencing results because no rare missense variant was identified in our controls).

Enrichment of Signals of Association at Coding Variants

The genetic-burden tests described above do not test the role of single variants whether the variants are of low frequency (0.1%–5%) or common (>5%) in the general population. To extend our sequencing study to independent modest-effect variants that are sufficiently frequent to be cataloged and genotyped with current arrays, we took advantage of two large RA case-control genotyping data sets: (1) 8,246 RA cases and 17,741 controls of European ancestry from seven collections and genotyped at high density across autoimmune-disease-related loci with the Illumina Immunochip (iChip) platform[26] (and unpublished data) and (2) a GWAS data set including 2,363 RA cases and 17,872 controls of European ancestry from four collections[28] (Table S3). The iChip data described here include a new set of 1,024 RA cases and 1,863 controls not previously analyzed. Also, our GWAS data set reported here was imputed genome-wide with haplotype-phased 1000 Genomes CEU data as a reference panel for the improvement of coverage of low-frequency variants.

We first assessed the coverage of variants discovered by sequencing in the iChip and GWAS data sets. By examining good proxies of the sequencing variants ($r^2 > 0.8$, as determined with the 1000 Genomes CEU data), we found that 42 out of 60 SNPs observed with a MAF > 0.5% in controls in our sequencing study were tagged in our quality-filtered genotyping data set. In contrast, 5 out of 211 SNPs with a MAF < 0.5% in controls were tagged in the genotyping data set (Figure 1). We focused on this set of 47 SNPs for association testing in our combined sample set of 10,609 RA $ACPA^+$ cases and 35,605 controls. Of the 47 coding variants, 23 were nonsynonymous variants and 24 were synonymous variants.

To test these 47 coding variants for association with RA risk, we performed a meta-analysis of all data combined and compared the association p values with meta-analysis results after 1,000 case-controls permutations of the 11 data sets. After excluding variants in LD with each other ($r^2 > 0.4$), we observed 16 (37%) of the 43 independent variants with p < 0.05 for association with RA, whereas three were expected by chance alone ($p_{enrichment} = 1.4 \times 10^{-8}$) (Table S4). The signal appeared to be driven by both nonsynonymous variants (9/22 variants with p < 0.05; $p_{enrichment} = 8.5 \times 10^{-4}$) and synonymous variants (7/21 variants with p < 0.05; $p_{enrichment} = 5.9 \times 10^{-4}$).

Conditional Analysis on Established, Common RA Risk Alleles

To assess whether the signal observed at coding variants was driven by the established common RA risk allele, we first evaluated the LD between the 16 associated coding variants and the known GWAS signals (Table S4). Three of the 16 coding variants were in strong LD with the known GWAS SNPs ($r^2 > 0.8$), including the well-known causal missense variant rs2476601 in *PTPN22*.[39–42] An additional four coding variants were in moderate LD with the known GWAS SNP (2/16 variants with $r^2 = 0.4$–0.8 and 2/16 variants with $r^2 = 0.1$–0.4), and the other nine coding variants had an $r^2 < 0.05$ with the GWAS SNP.

To account for the known GWAS signal, we performed conditional analyses adjusting for the best hit identified in the meta-analysis at each locus. After conditional analysis, we identified 10 of 43 (23%) coding variants showing p < 0.05 for association with RA risk; this was more than expected by chance alone ($p_{enrichment} = 6.4 \times 10^{-4}$). This signal was driven by both nonsynonymous (6/22 variants; $p_{enrichment} = 8.7 \times 10^{-3}$) and synonymous variants (4/21 variants; $p_{enrichment} = 0.025$) (Figure 3 and Table 2). One of the coding variants with p < 0.05 is *TNFAIP3* (MIM 191163), and a proxy of this nonsynonymous variant ($r^2 = 0.62$) has been previously implicated as an RA risk allele.[43] After removing this coding *TNFAIP3* variant, we still observed evidence of enrichment ($p_{enrichment} = 2 \times 10^{-3}$; Figure S2).

Although these observations strongly suggest that some of the coding variants analyzed here are true risk alleles for RA independent of the known RA risk alleles, no single variant survived a stringent correction for multiple hypotheses testing given the number of SNPs tested across each locus. However, we did observe that five out of the ten coding variants with $p_{condition} < 0.05$ were in LD ($r^2 > 0.7$) with the best hit in the conditional analysis ($p_{enrichment} = 1.7 \times 10^{-3}$); the five remaining variants had $r^2 < 0.05$ with the best hit (Table 2). That is, if the coding variants had no effect on risk of RA (i.e., spurious association), it is unlikely that the coding variants would be in high LD with the second strongest signal of association after conditioning on the known GWAS signal.

We also performed a conditional analysis adjusting for the coding variants and compared the association signal at the known GWAS hit in the meta-analysis and the conditional analysis (Table S5). As indicated by the consistent ORs in both analyses, the ten coding variants showing p < 0.05 when we adjusted for the known GWAS risk alleles did not contribute to the known GWAS signals of association.

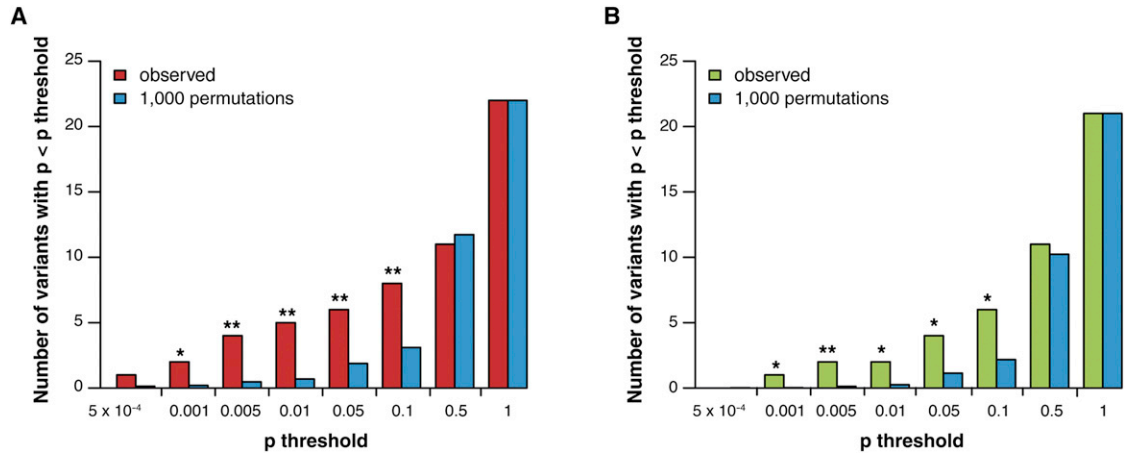


Figure 3. Enrichment of Nominal Association Signal Driven by Nonsynonymous and Synonymous Variants in the Conditional Analysis
The numbers of nonsynonymous variants (A) and synonymous variants (B) reaching the p < p threshold in our conditional analysis or after 1,000 permutations of the phenotypes are shown. Significant enrichment of SNPs with the p < p threshold in our conditional analysis was assessed by a Fisher's exact test (*p < 0.05 and **p < 0.01).

Together, these data argue for a model in which at least some RA risk loci harbor additional yet-unidentified low-frequency or common risk alleles with a modest effect size (OR ≈ 1.2; Table 2). Importantly, these yet-undiscovered independent association signals are most likely driven by variants within the protein-coding sequence of genes. Our results suggest that not only nonsynonymous but also synonymous variants play a role in these independent signals of association.

Independent Signal of Association at *CD2*

Most of the ten candidate coding variants identified above showed either a small effect size or a low frequency, which did not allow us to perform detailed follow-up analyses with enough power considering the sample size of our genotyping data set. However, one common missense variant in *CD2* (MIM 186990), rs699738 (c.798C>A [p.His266Gln]), showed a $p_{condition} = 0.0017$ and was in complete LD ($r^2 = 1$) with the best signal of association in the conditional analysis (MAF = 0.09 and $OR_{condition} = 0.88$). We thus performed more detailed haplotype and conditional analyses at the *CD2* locus. Our meta-analysis and conditional analysis provided suggestive statistical evidence of association at both a common noncoding variant (rs624988 [MAF = 0.4]) and the independent missense variant rs699738 (Table 2 and Figures 4A and 4B). Although rs699738 was not directly analyzed in our meta-analysis, we identified proxy SNPs that were in complete LD ($r^2 = 1$) and that were included in our meta-analysis (these were rs798036, rs798037, rs798044, and rs810048). As shown in Figure 4C, these proxy SNPs showed consistent effect sizes among the 11 data sets and thus association with RA risk in the meta-analysis. After adjusting for both rs624988 and rs798036 in a conditional analysis, we observed no signal at p < 0.01 (Figure S3).

In a haplotype analysis using iChip data (Figure 4D), we observed that the perfect proxy of the missense variant (rs798036) was independent of the common variant driving the best signal of association at the locus (rs624988). For the three common haplotypes formed by these two *CD2* variants, there was a dose-dependent effect on risk of RA on the basis of point estimates: the A-T haplotype with both risk alleles had a higher susceptibility (OR = 1.22) than did the G-A haplotype with both nonrisk alleles, whereas the G-T haplotype carrying the *CD2* missense variant alone demonstrated an intermediate effect (OR = 1.14). (The A-A haplotype carrying the noncoding risk variant alone was infrequent [1.6%] in the general population and had an OR of 1.45 and a 95% confidence interval of 1.13–1.85). Consistent with Figure 4B, we observed a significant association for the G-A haplotype carrying the *CD2* missense variant risk allele when controlling for the noncoding variant rs624988 (p = 0.015). We further assessed the significance of this result by performing 5,000 case-control permutations while preserving genotypes for the common variant (fixing case-control allele frequencies and ORs) and found that the observation of a signal of association at $p \leq 0.015$ after controlling for the effect of rs624988 is beyond what might be expected by chance alone (p = 0.014). Considering both variants and all haplotypes combined, we observed a highly significant ($p = 4.6 \times 10^{-6}$) association between the *CD2* locus and risk of RA.

Although the *CD2* missense variant represents a strong candidate for one of the causal alleles at the *CD2* locus, we considered the possibility that a noncoding variant could also be responsible for the independent signal at *CD2*. We identified 16 SNPs in strong LD ($r^2 > 0.8$) with missense variant rs699738. These SNPs included nine intronic variants and seven variants in the 3′ region of the coding sequence. We used several publically available tools and data sets to annotate the missense variant (rs699738) and the 16 noncoding variants (Table S6): (1) PolyPhen[21] and SIFT[33] to predict the function of the

Table 2. Coding Variants Showing Nominal Signal of Association in the Conditional Analysis

Gene	Coding Variant					Best Hit$_{condition}$[a]		
	Reference SNP ID	MAF	OR$_{meta}$ (95% CI)	p$_{meta}$	p$_{condition}$	Reference SNP ID	p$_{condition}$	r^{2}[b]
Missense Variants								
TNFAIP3	rs2230926	0.038	1.38 (1.30–1.46)	6.8×10^{-14}	1.4×10^{-9}	rs58721818	3.3×10^{-10}	**0.87**
FCGR2A	rs1801274	0.494	1.10 (1.08–1.12)	2.4×10^{-7}	5×10^{-4}	index	5×10^{-4}	**1**
CD2	rs699738[c]	0.086	0.88 (0.82–0.94)	1.2×10^{-4}	0.0017	index	0.0017	**1**
TNFRSF14	rs2234163[c]	0.006	1.52 (1.28–1.76)	7.6×10^{-4}	0.0019	rs1886731	8×10^{-4}	~0
BLK	rs55758736	0.012	0.79 (0.61–0.97)	0.01	0.0077	rs77072957	1×10^{-4}	~0
TAGAP	rs41267765	0.029	1.18 (1.08–1.20)	0.0018	0.0107	rs112904761	0.0017	**0.76**
Synonymous Variants								
IL2RA	rs2228150	0.03	1.25 (1.15–1.35)	6.6×10^{-6}	6×10^{-4}	rs11256360	5×10^{-4}	**0.88**
IL2RB	rs228953	0.44	0.95 (0.93–0.97)	0.0023	0.0042	rs5756391	4×10^{-4}	~0
CD58	rs35768283	0.02	1.20 (1.08–1.32)	0.0043	0.0117	rs798036	0.0017	~0
TRAF1	rs3747841[c]	0.013	0.89 (0.72–1.07)	0.17	0.03	rs76418192	0.0037	~0

The following abbreviations are used: MAF, minor allele frequency; OR, odds ratio; and CI, confidence interval.
[a]Best signal of association in the conditional analysis.
[b]LD between the coding variant and the best signal of association. Boldface indicates $r^2 > 0.7$.
[c]Coding variants represented by a proxy in complete LD in the meta-analysis.

missense variant, (2) GERP[34] to identify evolutionary constrained sites, and (3) eQTL data in human CD4$^+$ T cells and monocytes.[44,45] Missense variant rs699738 induces the substitution of a histidine for a glutamine at position 266 of CD2 (p.His266Gln) but was predicted to be benign by both PolyPhen and SIFT. None of the 17 noncoding variants showed a high conservation score in mammals, and we found no evidence that these SNPs are associated with *CD2* expression as eQTLs. All together, we found no evidence to support a noncoding variant as being potentially damaging.

Discussion

The extent to which variants across the allele-frequency spectrum contribute to complex traits such as risk of RA and the actual contribution of variants within protein-coding sequences are two important issues in human genetics. In this study, we addressed these two questions by deep exon-sequencing and large-scale genotyping across 25 genes from RA risk loci discovered by GWASs. Our findings support a model in which most GWAS findings in RA are due to common variants that fall outside of the protein-coding sequences of genes. More importantly, however, we provide evidence of independent association signals driven by coding variants within a subset of biological candidate genes identified by GWASs and show that rare, low-frequency, and common variants with a small to moderate effect size participate in the missing genetic contribution to RA.

There are several important implications of our study. First, we assessed the contribution of coding variants to risk of RA. Although most of the candidate genes that we sequenced did not have evidence of rare protein-coding mutations contributing to risk of RA, we did find that, in RA cases, two genes (*IL2RA* and *IL2RB*) harbor an accumulation of rare missense variants that result in a moderate burden signal of association ($p = 0.007$ and $p = 0.018$, respectively; Figure 2). Further, by testing the aggregate signal of association at variants that have a MAF $> 0.1\%$ and that lie within protein-coding sequences from a subset of biological candidate genes, we have demonstrated a significant enrichment of coding variants with $p < 0.05$ ($p_{enrichment} = 6.4 \times 10^{-4}$; Table 2 and Figure 3) and have shown that these associated coding variants are observed in LD ($r^2 > 0.7$) with the second best hit at the loci more frequently than would be expected by chance alone ($p_{enrichment} = 1.7 \times 10^{-3}$). Although this analysis cannot definitively confirm which coding variants are causative, our study does suggest that many will ultimately contribute to risk of RA. To illustrate these findings, we provide further evidence that one of these associated coding variants, altering the sequence of *CD2*, is independent of the common noncoding variant driving the best signal of association at the locus, and we have shown that these two variants form three common haplotypes conferring risk of RA in a dose-dependent manner (Figure 4).

A second implication involves protein-coding mutations to help identify the disease-causative genes at GWAS loci, which in turn provides insight into disease biology.

As discussed above, this study identified three genes carrying missense variants associated with risk of RA: *IL2RA*, *IL2RB*, and *CD2*. *IL2RA* encodes IL-2Rα (CD25), and *IL2RB* encodes IL-2Rβ (CD122); all together, CD25,

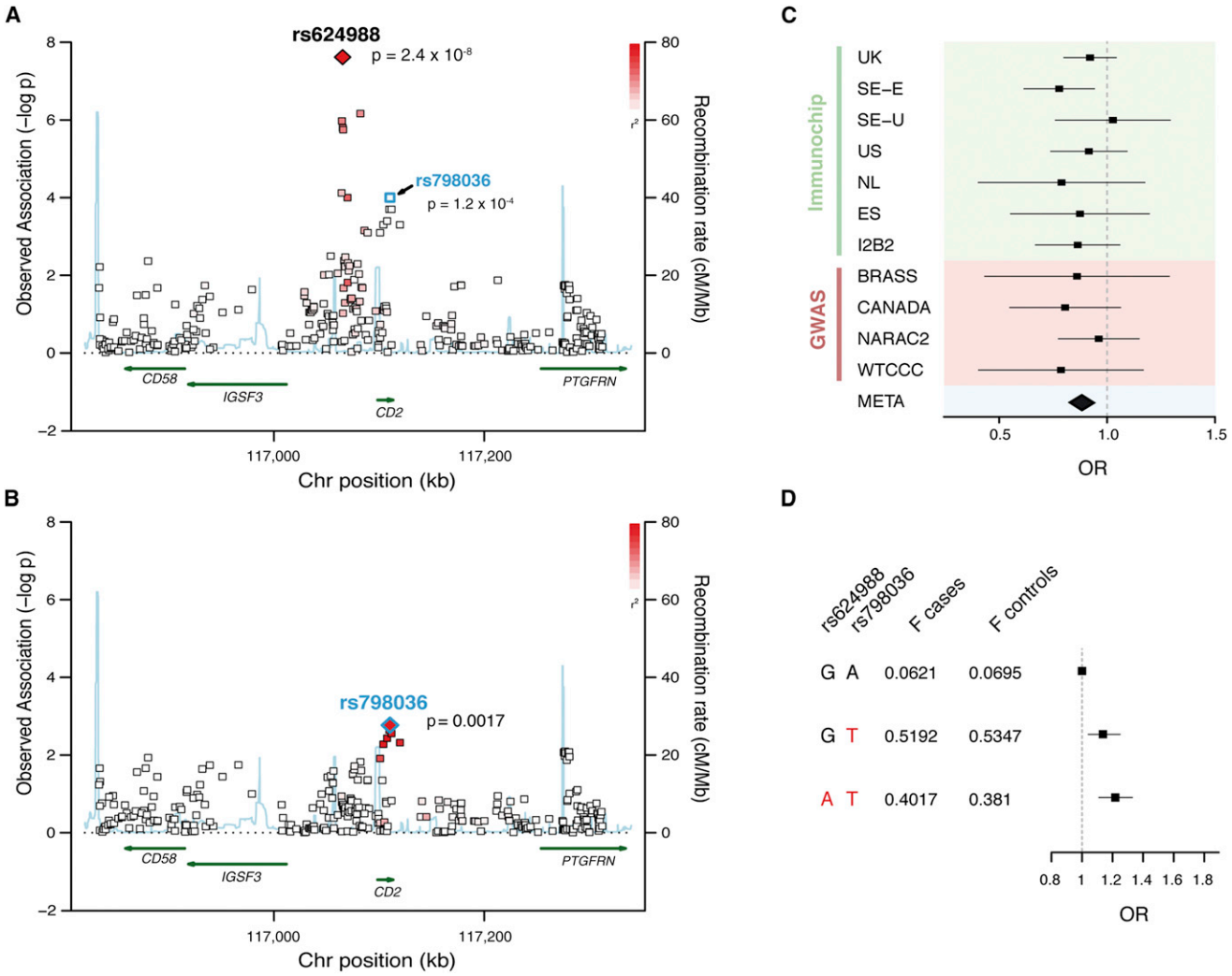


Figure 4. Evidence of an Independent Signal of Association at the *CD2* Locus
(A and B) Association results from the meta-analysis (A) and the conditional analysis (B). In these analyses, missense SNP rs699738 is represented by a group of SNPs, including rs798036 (highlighted in blue), in perfect LD ($r^2 = 1$). In each analysis, the best signal of association is indicated by a diamond. Only SNPs present in more than five collections are shown.
(C) ORs and 95% confidence interval in the independent cohorts and the meta-analysis.
(D) Results from the haplotype analysis using the best signal in the meta-analysis (rs624988) and rs798036. In this analysis, only genotype data were used (7,222 RA cases and 15,870 controls). With this subset of samples, rs624988 and rs798036 reached $p = 1.3 \times 10^{-5}$ and $p = 2 \times 10^{-3}$ in the meta-analysis, respectively. The overall *CD2* variation due to rs624988 and rs798036 contributed to RA with $p = 4 \times 10^{-6}$. The RA risk allele is highlighted in red.

CD122, and the common gamma chain, CD132, constitute the three subunits of the high-affinity IL-2 receptor (IL2R).[46] Studies have reported that the inactivation of either CD25 or CD122 in mice results in lethal autoimmunity.[47,48] Furthermore, a risk allele in *IL2RA* has been linked with a decreased function of T regulatory (Treg) cells in type 1 diabetes.[49,50] *CD2* encodes a cell-surface antigen expressed on T cells. CD2 coactivation has been shown to induce the suppression of T cell proliferation by activation of $CD4^+$ $CD25^{hi}$ Treg cells.[51,52] Interestingly, several reports have shown that Treg cells, which control proinflammatory responses, are functionally compromised in individuals with RA.[53–55] Further validation and functional analyses will be needed for supporting this observation and for assessing whether the protein-coding variants within these three genes affect the IL-2-signaling pathway, the differentiation or activity of Treg cells, or other disease-related processes to be determined.

A third implication addresses the allele-frequency spectrum and effect size of the variants participating in the risk of RA. We have previously used simulations to show that rare, low-frequency, and common variants of small to modest effects participate in the missing genetic contribution to RA.[6] Here, we provide data to empirically support this model. First, the moderate burden signals of association at *IL2RA* and *IL2RB* implicate rare variants as participating in the risk of disease. Second, the significant enrichment of independent coding variants associated with RA highlights low-frequency and common variants with a modest effect size (OR ≈ 1.2; Table 2) and shows

that these signals of association are independent of the known RA risk signals identified by GWASs (Table 2 and Table S5). Importantly, these conclusions are based on a small subset of candidate genes selected for sequencing. A more comprehensive approach to genome sequencing in RA will be required for assessing whether these findings can be extrapolated to the remainder of the genome. Ultimately, sequence data on the entire genome in large sample collections will be required for understanding the complete genetic architecture of RA risk.

A fourth implication pertains to the design and interpretation of future large-scale sequencing and genotyping studies in RA, as well as potentially other complex traits. Our results strongly suggest that genes implicated by GWASs serve as excellent candidates for future sequencing studies. For example, it might be informative to use computational methods such as GRAIL to prioritize genes for sequencing or even the interpretation of results that emerge from whole-genome-sequencing studies. Our results also underscore the importance of conditional analysis for adjusting for the best signal of association in the GWAS region. Lastly, our study emphasizes the need for very large sample sizes for teasing apart independent signals (Figures S4 and S5). For rare variants, we estimate that a sample size of at least 3,300 RA cases and 3,300 controls would be required for reaching a significant threshold of 2.5×10^{-6} (which corresponds to $p = 0.05$ corrected for 20,000 independent tests or genes) given the effect size observed for *IL2RA*, whereas >5,500 RA cases and 5,500 controls would be required for reaching $p = 2.5 \times 10^{-6}$ at *IL2RB* (Figure S4). Our findings are consistent with the published literature on genetic-burden tests for rare variants; no candidate-gene study reported to date has reported an overwhelming signal of statistical significance.[35]

There are important limitations of our study. First, we use pooled sequencing data to estimate allele frequency in our tests of rare variants. Our sensitivity and specificity analysis (Table S2 and Figure S1) suggest that this strategy was effective at discovering rare variants in that it had little evidence of false-positive findings due to technical artifacts. Second, we did not attempt to validate the rare variants identified in *IL2RA* and *IL2RB* by alternative methods (like Sanger sequencing) but only used an indirect strategy to assess the sensitivity and specificity of our sequencing results by using GWAS data from 250 samples that were included in our targeted sequencing. The results of this analysis, together with the very high coverage in our sequencing study and the stringent filters applied in variant calling, support the quality of our sequencing results. Third and finally, we used common variants to match cases and controls in our pooled sequencing study. It is possible that common variants do not adequately capture the underlying population structure of rare variants,[56] which might have biased the association statistics in our genetic burden tests.

In conclusion, our study provides evidence of independent RA risk alleles driven by variants in the protein-coding sequence of genes discovered by GWASs. These alleles are rare, of low frequency, and common in the general population, and each contributes a small effect to disease risk. Our findings suggest that integrating sequence data with large-scale genotyping will serve as an effective strategy for discovering RA risk alleles in the future.

Supplemental Data

Supplemental Data include five figures and six tables and can be found with this article online at http://www.cell.com/AJHG.

Acknowledgments

R.M.P. is supported by National Institutes of Health (NIH) grants R01-AR057108, R01-AR056768, U01-GM092691, and R01-AR059648 and holds a Career Award for Medical Scientists from the Burroughs Wellcome Fund. F.K. was supported by a European Union international outgoing fellowship. K.P.L. is supported by NIH K08AR060257. Immunochip genotype data from the i2b2 samples and electronic-medical-record dataset were funded in part by NIH grant U54-LM008748. K.A.S. holds the Sherman Family Chair in Genomic Medicine and a Tier 1 Canada Research Chair. L.K. is supported by a grant from the European Research Council. L.P. is supported by Swedish Research Council grant 521-2009-3185.

Received: July 27, 2012
Revised: September 4, 2012
Accepted: November 26, 2012
Published: December 20, 2012

Web Resources

The URLs for data presented herein are as follows:

Genetic Power Calculator, http://pngu.mgh.harvard.edu/~purcell/gpc/
Online Mendelian Inheritance in Man (OMIM), http://www.omim.org
NHLBI Exome Sequencing Project Exome Variant Server, http://evs.gs.washington.edu/EVS/
University of Chicago's eQTL data repository, http://eqtl.uchicago.edu/cgi-bin/gbrowse/eqtl/

References

1. Manolio, T.A., Collins, F.S., Cox, N.J., Goldstein, D.B., Hindorff, L.A., Hunter, D.J., McCarthy, M.I., Ramos, E.M., Cardon, L.R., Chakravarti, A., et al. (2009). Finding the missing heritability of complex diseases. Nature *461*, 747–753.
2. Kryukov, G.V., Pennacchio, L.A., and Sunyaev, S.R. (2007). Most rare missense alleles are deleterious in humans: Implications for complex disease and association studies. Am. J. Hum. Genet. *80*, 727–739.
3. Marth, G.T., Yu, F., Indap, A.R., Garimella, K., Gravel, S., Leong, W.F., Tyler-Smith, C., Bainbridge, M., Blackwell, T., Zheng-Bradley, X., et al.; 1000 Genomes Project (2011). The functional spectrum of low-frequency coding variation. Genome Biol. *12*, R84.
4. Purcell, S.M., Wray, N.R., Stone, J.L., Visscher, P.M., O'Donovan, M.C., Sullivan, P.F., and Sklar, P.; International

Schizophrenia Consortium (2009). Common polygenic variation contributes to risk of schizophrenia and bipolar disorder. Nature *460*, 748–752.

5. Yang, J., Benyamin, B., McEvoy, B.P., Gordon, S., Henders, A.K., Nyholt, D.R., Madden, P.A., Heath, A.C., Martin, N.G., Montgomery, G.W., et al. (2010). Common SNPs explain a large proportion of the heritability for human height. Nat. Genet. *42*, 565–569.
6. Stahl, E.A., Wegmann, D., Trynka, G., Gutierrez-Achury, J., Do, R., Voight, B.F., Kraft, P., Chen, R., Kallberg, H.J., Kurreeman, F.A., et al.; Diabetes Genetics Replication and Meta-analysis Consortium; Myocardial Infarction Genetics Consortium (2012). Bayesian inference analyses of the polygenic architecture of rheumatoid arthritis. Nat. Genet. *44*, 483–489.
7. Cohen, J.C., Boerwinkle, E., Mosley, T.H., Jr., and Hobbs, H.H. (2006). Sequence variations in PCSK9, low LDL, and protection against coronary heart disease. N. Engl. J. Med. *354*, 1264–1272.
8. Nejentsev, S., Walker, N., Riches, D., Egholm, M., and Todd, J.A. (2009). Rare variants of IFIH1, a gene implicated in antiviral responses, protect against type 1 diabetes. Science *324*, 387–389.
9. Johansen, C.T., Wang, J., McIntyre, A.D., Martins, R.A., Ban, M.R., Lanktree, M.B., Huff, M.W., Péterfy, M., Mehrabian, M., Lusis, A.J., et al. (2012). Excess of rare variants in non-genome-wide association study candidate genes in patients with hypertriglyceridemia. Circ. Cardiovasc. Genet. *5*, 66–72.
10. Momozawa, Y., Mni, M., Nakamura, K., Coppieters, W., Almer, S., Amininejad, L., Cleynen, I., Colombel, J.F., de Rijk, P., Dewit, O., et al. (2011). Resequencing of positional candidates identifies low frequency IL23R coding variants protecting against inflammatory bowel disease. Nat. Genet. *43*, 43–47.
11. Rivas, M.A., Beaudoin, M., Gardet, A., Stevens, C., Sharma, Y., Zhang, C.K., Boucher, G., Ripke, S., Ellinghaus, D., Burtt, N., et al.; National Institute of Diabetes and Digestive Kidney Diseases Inflammatory Bowel Disease Genetics Consortium (NIDDK IBDGC); United Kingdom Inflammatory Bowel Disease Genetics Consortium; International Inflammatory Bowel Disease Genetics Consortium (2011). Deep resequencing of GWAS loci identifies independent rare variants associated with inflammatory bowel disease. Nat. Genet. *43*, 1066–1073.
12. Raychaudhuri, S., Iartchouk, O., Chin, K., Tan, P.L., Tai, A.K., Ripke, S., Gowrisankar, S., Vemuri, S., Montgomery, K., Yu, Y., et al. (2011). A rare penetrant mutation in CFH confers high risk of age-related macular degeneration. Nat. Genet. *43*, 1232–1236.
13. Bonnefond, A., Clément, N., Fawcett, K., Yengo, L., Vaillant, E., Guillaume, J.L., Dechaume, A., Payne, F., Roussel, R., Czernichow, S., et al.; Meta-Analysis of Glucose and Insulin-Related Traits Consortium (MAGIC) (2012). Rare MTNR1B variants impairing melatonin receptor 1B function contribute to type 2 diabetes. Nat. Genet. *44*, 297–301.
14. Leblond, C.S., Heinrich, J., Delorme, R., Proepper, C., Betancur, C., Huguet, G., Konyukh, M., Chaste, P., Ey, E., Rastam, M., et al. (2012). Genetic and functional analyses of SHANK2 mutations suggest a multiple hit model of autism spectrum disorders. PLoS Genet. *8*, e1002521.
15. Jordan, C.T., Cao, L., Roberson, E.D., Duan, S., Helms, C.A., Nair, R.P., Duffin, K.C., Stuart, P.E., Goldgar, D., Hayashi, G., et al. (2012). Rare and common variants in CARD14, encoding an epidermal regulator of NF-kappaB, in psoriasis. Am. J. Hum. Genet. *90*, 796–808.
16. Klareskog, L., Catrina, A.I., and Paget, S. (2009). Rheumatoid arthritis. Lancet *373*, 659–672.
17. Padyukov, L., Silva, C., Stolt, P., Alfredsson, L., and Klareskog, L. (2004). A gene-environment interaction between smoking and shared epitope genes in HLA-DR provides a high risk of seropositive rheumatoid arthritis. Arthritis Rheum. *50*, 3085–3092.
18. Kurreeman, F., Liao, K., Chibnik, L., Hickey, B., Stahl, E., Gainer, V., Li, G., Bry, L., Mahan, S., Ardlie, K., et al. (2011). Genetic basis of autoantibody positive and negative rheumatoid arthritis risk in a multi-ethnic cohort derived from electronic health records. Am. J. Hum. Genet. *88*, 57–69.
19. Raychaudhuri, S., Plenge, R.M., Rossin, E.J., Ng, A.C., Purcell, S.M., Sklar, P., Scolnick, E.M., Xavier, R.J., Altshuler, D., and Daly, M.J.; International Schizophrenia Consortium (2009). Identifying relationships among genomic disease regions: Predicting genes at pathogenic SNP associations and rare deletions. PLoS Genet. *5*, e1000534.
20. Li, H., Ruan, J., and Durbin, R. (2008). Mapping short DNA sequencing reads and calling variants using mapping quality scores. Genome Res. *18*, 1851–1858.
21. Adzhubei, I.A., Schmidt, S., Peshkin, L., Ramensky, V.E., Gerasimova, A., Bork, P., Kondrashov, A.S., and Sunyaev, S.R. (2010). A method and server for predicting damaging missense mutations. Nat. Methods *7*, 248–249.
22. Madsen, B.E., and Browning, S.R. (2009). A groupwise association test for rare mutations using a weighted sum statistic. PLoS Genet. *5*, e1000384.
23. Price, A.L., Kryukov, G.V., de Bakker, P.I., Purcell, S.M., Staples, J., Wei, L.J., and Sunyaev, S.R. (2010). Pooled association tests for rare variants in exon-resequencing studies. Am. J. Hum. Genet. *86*, 832–838.
24. Morris, A.P., and Zeggini, E. (2010). An evaluation of statistical approaches to rare variant analysis in genetic association studies. Genet. Epidemiol. *34*, 188–193.
25. Neale, B.M., Rivas, M.A., Voight, B.F., Altshuler, D., Devlin, B., Orho-Melander, M., Kathiresan, S., Purcell, S.M., Roeder, K., and Daly, M.J. (2011). Testing for an unusual distribution of rare variants. PLoS Genet. *7*, e1001322.
26. Eyre, S., Bowes, J., Diogo, D., Lee, A., Barton, A., Martin, P., Zhernakova, A., Stahl, E., Viatte, S., McAllister, K., et al.; Biologics in Rheumatoid Arthritis Genetics and Genomics Study Syndicate; Wellcome Trust Case Control Consortium (2012). High-density genetic mapping identifies new susceptibility loci for rheumatoid arthritis. Nat. Genet. *44*, 1336–1340.
27. Kremer, J. (2005). The CORRONA database. Ann. Rheum. Dis. *64*(*Suppl 4*), iv37–iv41.
28. Stahl, E.A., Raychaudhuri, S., Remmers, E.F., Xie, G., Eyre, S., Thomson, B.P., Li, Y., Kurreeman, F.A., Zhernakova, A., Hinks, A., et al.; BIRAC Consortium; YEAR Consortium (2010). Genome-wide association study meta-analysis identifies seven new rheumatoid arthritis risk loci. Nat. Genet. *42*, 508–514.
29. Liao, K.P., Cai, T., Gainer, V., Goryachev, S., Zeng-treitler, Q., Raychaudhuri, S., Szolovits, P., Churchill, S., Murphy, S., Kohane, I., et al. (2010). Electronic medical records for discovery research in rheumatoid arthritis. Arthritis Care Res (Hoboken) *62*, 1120–1127.
30. Price, A.L., Patterson, N.J., Plenge, R.M., Weinblatt, M.E., Shadick, N.A., and Reich, D. (2006). Principal components

analysis corrects for stratification in genome-wide association studies. Nat. Genet. *38*, 904–909.
31. Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M.A., Bender, D., Maller, J., Sklar, P., de Bakker, P.I., Daly, M.J., and Sham, P.C. (2007). PLINK: A tool set for whole-genome association and population-based linkage analyses. Am. J. Hum. Genet. *81*, 559–575.
32. de Bakker, P.I., Ferreira, M.A., Jia, X., Neale, B.M., Raychaudhuri, S., and Voight, B.F. (2008). Practical aspects of imputation-driven meta-analysis of genome-wide association studies. Hum. Mol. Genet. *17*(R2), R122–R128.
33. Kumar, P., Henikoff, S., and Ng, P.C. (2009). Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm. Nat. Protoc. *4*, 1073–1081.
34. Cooper, G.M., Stone, E.A., Asimenos, G., Green, E.D., Batzoglou, S., and Sidow, A.; NISC Comparative Sequencing Program (2005). Distribution and intensity of constraint in mammalian genomic sequence. Genome Res. *15*, 901–913.
35. Kiezun, A., Garimella, K., Do, R., Stitziel, N.O., Neale, B.M., McLaren, P.J., Gupta, N., Sklar, P., Sullivan, P.F., Moran, J.L., et al. (2012). Exome sequencing and the genetic basis of complex traits. Nat. Genet. *44*, 623–630.
36. Purcell, S., Cherny, S.S., and Sham, P.C. (2003). Genetic Power Calculator: Design of linkage and association genetic mapping studies of complex traits. Bioinformatics *19*, 149–150.
37. Rossin, E.J., Lage, K., Raychaudhuri, S., Xavier, R.J., Tatar, D., Benita, Y., Cotsapas, C., and Daly, M.J.; International Inflammatory Bowel Disease Genetics Constortium (2011). Proteins encoded in genomic regions associated with immune-mediated disease physically interact and suggest underlying biology. PLoS Genet. *7*, e1001273.
38. Chen, R., Stahl, E.A., Kurreeman, F.A., Gregersen, P.K., Siminovitch, K.A., Worthington, J., Padyukov, L., Raychaudhuri, S., and Plenge, R.M. (2011). Fine mapping the TAGAP risk locus in rheumatoid arthritis. Genes Immun. *12*, 314–318.
39. Arechiga, A.F., Habib, T., He, Y., Zhang, X., Zhang, Z.Y., Funk, A., and Buckner, J.H. (2009). Cutting edge: The PTPN22 allelic variant associated with autoimmunity impairs B cell signaling. J. Immunol. *182*, 3343–3347.
40. Rieck, M., Arechiga, A., Onengut-Gumuscu, S., Greenbaum, C., Concannon, P., and Buckner, J.H. (2007). Genetic variation in PTPN22 corresponds to altered function of T and B lymphocytes. J. Immunol. *179*, 4704–4710.
41. Vang, T., Congia, M., Macis, M.D., Musumeci, L., Orrú, V., Zavattari, P., Nika, K., Tautz, L., Taskén, K., Cucca, F., et al. (2005). Autoimmune-associated lymphoid tyrosine phosphatase is a gain-of-function variant. Nat. Genet. *37*, 1317–1319.
42. Zhang, J., Zahir, N., Jiang, Q., Miliotis, H., Heyraud, S., Meng, X., Dong, B., Xie, G., Qiu, F., Hao, Z., et al. (2011). The autoimmune disease-associated PTPN22 variant promotes calpain-mediated Lyp/Pep degradation associated with lymphocyte and dendritic cell hyperresponsiveness. Nat. Genet. *43*, 902–907.
43. Orozco, G., Hinks, A., Eyre, S., Ke, X., Gibbons, L.J., Bowes, J., Flynn, E., Martin, P., Wilson, A.G., Bax, D.E., et al.; Wellcome Trust Case Control Consortium; YEAR consortium (2009). Combined effects of three independent SNPs greatly increase the risk estimate for RA at 6q23. Hum. Mol. Genet. *18*, 2693–2699.
44. Dimas, A.S., Deutsch, S., Stranger, B.E., Montgomery, S.B., Borel, C., Attar-Cohen, H., Ingle, C., Beazley, C., Gutierrez Arcelus, M., Sekowska, M., et al. (2009). Common regulatory variation impacts gene expression in a cell type-dependent manner. Science *325*, 1246–1250.
45. Zeller, T., Wild, P., Szymczak, S., Rotival, M., Schillert, A., Castagne, R., Maouche, S., Germain, M., Lackner, K., Rossmann, H., et al. (2010). Genetics and beyond—The transcriptome of human monocytes and disease susceptibility. PLoS ONE *5*, e10693.
46. Malek, T.R. (2008). The biology of interleukin-2. Annu. Rev. Immunol. *26*, 453–479.
47. Suzuki, H., Kündig, T.M., Furlonger, C., Wakeham, A., Timms, E., Matsuyama, T., Schmits, R., Simard, J.J., Ohashi, P.S., Griesser, H., et al. (1995). Deregulated T cell activation and autoimmunity in mice lacking interleukin-2 receptor beta. Science *268*, 1472–1476.
48. Willerford, D.M., Chen, J., Ferry, J.A., Davidson, L., Ma, A., and Alt, F.W. (1995). Interleukin-2 receptor alpha chain regulates the size and content of the peripheral lymphoid compartment. Immunity *3*, 521–530.
49. Dendrou, C.A., Plagnol, V., Fung, E., Yang, J.H., Downes, K., Cooper, J.D., Nutland, S., Coleman, G., Himsworth, M., Hardy, M., et al. (2009). Cell-specific protein phenotypes for the autoimmune locus IL2RA using a genotype-selectable human bioresource. Nat. Genet. *41*, 1011–1015.
50. Garg, G., Tyler, J.R., Yang, J.H., Cutler, A.J., Downes, K., Pekalski, M., Bell, G.L., Nutland, S., Peakman, M., Todd, J.A., et al. (2012). Type 1 diabetes-associated IL2RA variation lowers IL-2 signaling and contributes to diminished CD4+ CD25+ regulatory T cell function. J. Immunol. *188*, 4644–4653.
51. Baecher-Allan, C.M., Costantino, C.M., Cvetanovich, G.L., Ashley, C.W., Beriou, G., Dominguez-Villar, M., and Hafler, D.A. (2011). CD2 costimulation reveals defective activity by human CD4+CD25(hi) regulatory cells in patients with multiple sclerosis. J. Immunol. *186*, 3317–3326.
52. De Jager, P.L., Baecher-Allan, C., Maier, L.M., Arthur, A.T., Ottoboni, L., Barcellos, L., McCauley, J.L., Sawcer, S., Goris, A., Saarela, J., et al. (2009). The role of the CD58 locus in multiple sclerosis. Proc. Natl. Acad. Sci. USA *106*, 5264–5269.
53. Ehrenstein, M.R., Evans, J.G., Singh, A., Moore, S., Warnes, G., Isenberg, D.A., and Mauri, C. (2004). Compromised function of regulatory T cells in rheumatoid arthritis and reversal by anti-TNFalpha therapy. J. Exp. Med. *200*, 277–285.
54. Samson, M., Audia, S., Janikashvili, N., Ciudad, M., Trad, M., Fraszczak, J., Ornetti, P., Maillefert, J.F., Miossec, P., and Bonnotte, B. (2012). Brief report: Inhibition of interleukin-6 function corrects Th17/Treg cell imbalance in patients with rheumatoid arthritis. Arthritis Rheum. *64*, 2499–2503.
55. Wang, W., Shao, S., Jiao, Z., Guo, M., Xu, H., and Wang, S. (2012). The Th17/Treg imbalance and cytokine environment in peripheral blood of patients with rheumatoid arthritis. Rheumatol. Int. *32*, 887–893.
56. Mathieson, I., and McVean, G. (2012). Differential confounding of rare and common variants in spatially structured populations. Nat. Genet. *44*, 243–246.